A bibliometric application of clustering with relational constraint

Vladimir Batagelj
IMFM Ljubljana, IAM UP Koper and NRU HSE Moscow

16th Applied Statistics 2019
Ribno (Bled), Slovenia, September 22 - 25, 2019

Outline

Networks

Clustering with relational constraint

Example: Leader strategy

References

1 Networks
2 Clustering with relational constraint
3 Example: Leader strategy
4 References

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Current version of slides (September 24, 2019 at 00 : 20): slides PDF
https://github.com/bavla/cluRC/blob/master/doc/leader.pdf

Networks

A *network* is based on two sets – a set of *nodes* (vertices), that represent the selected *units*, and a set of *links* (lines), that represent *ties* between units. They determine a *graph*. A link can be *directed* – an *arc*, or *undirected* – an *edge*.
Additional data about nodes or links may be known – their *properties* (attributes). For example: name/label, type, age, value, ...

Network = Graph + Data

A *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of:

- a *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{A}$ is the set of arcs, $\mathcal{E}$ is the set of edges, and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$ is the set of links.
  $n = |\mathcal{V}|$, $m = |\mathcal{L}|$
- $\mathcal{P}$ *node value functions* / properties: $p : \mathcal{V} \rightarrow A$
- $\mathcal{W}$ *link value functions* / weights: $w : \mathcal{L} \rightarrow B$

Clustering with relational constraint

We shall deal with the *clustering problem* (Φ, P):
Determine the clustering $\mathbf{C}^\star \in \Phi$ for which

$$P(\mathbf{C}^\star) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where $\mathbf{U}$ is a finite *set of units*; C is a *cluster*, $\emptyset \subset C \subseteq \mathbf{U}$; $\mathbf{C} = \{C_i\}$ is a *clustering*; Φ is a set of *feasible clusterings*; and $P : \Phi \to \mathbb{R}_0^+$ is a *criterion function*.

The criterion function $P(\mathbf{C})$ combines "partial/local errors" into a "total error". Usually it takes the form: $P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C)$. The cluster-error $p(C)$ is usually expressed using a *dissimilarity* $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_0^+$.

Agglomerative method for relational constraints

Suppose that the units are described by attribute data a: $\mathbf{U} \rightarrow [\mathbf{U}]$ and are related by a binary *relation* $R \subseteq \mathbf{U} \times \mathbf{U}$ that determine the *relational data* or *network* $(\mathbf{U}, R, a)$ [Batagelj and Ferligoj(2000)].

We want to cluster the units according to some (dis)similarity of their descriptions, but also considering the relation R which imposes *constraints* on the set of feasible clusterings [Ferligoj and Batagelj(1982)], [Ferligoj and Batagelj(1983)], usually in the following form:

$$\Phi(R) = \{\mathbf{C} \in P(\mathbf{U}) : \text{each cluster } C \in \mathbf{C} \text{ induces a subgraph } (C, R \cap C \times C) \text{ in the graph } (\mathbf{U}, R) \text{ of the required type of connectedness}\}$$

Example: regionalization problem – group given territorial units into regions such that units inside the region will be similar and form contiguous part of the territory.

For the same relation R we can define different types of sets of feasible clusterings.

Types of connectedness

Some examples of *types of relational constraints*, $\Phi^i(R)$, are [Ferligoj and Batagelj(1983)]

clusterings	type of connectedness
$\Phi^1(R)$	weakly connected units
$\Phi^2(R)$	weakly connected units that contain at most one center
$\Phi^3(R)$	strongly connected units
$\Phi^4(R)$	clique
$\Phi^5(R)$	existence of a trail containing all the units of the cluster

In a directed graph a *trail* is a walk in which all arcs are distinct.
The set $R(X) = \{Y : X\,R\,Y\}$ is a *set of successors* of unit $X \in \mathbf{U}$ and, for a cluster $C \subseteq \mathbf{U}$, $R(C) = \bigcup_{X \in C} R(X)$. A set of units, $L \subseteq C$ is a *center* of a cluster C in the clustering of type $\Phi^2(R)$ iff the subgraph induced by L is strongly connected and $R(L) \cap (C \setminus L) = \emptyset$.

We can use both hierarchical and local optimization methods for solving some types of problems with relational constraint [Ferligoj and Batagelj(1982), Ferligoj and Batagelj(1983), Batagelj et al.(2014)].

Hierarchical method

Here, we present only the hierarchical method:

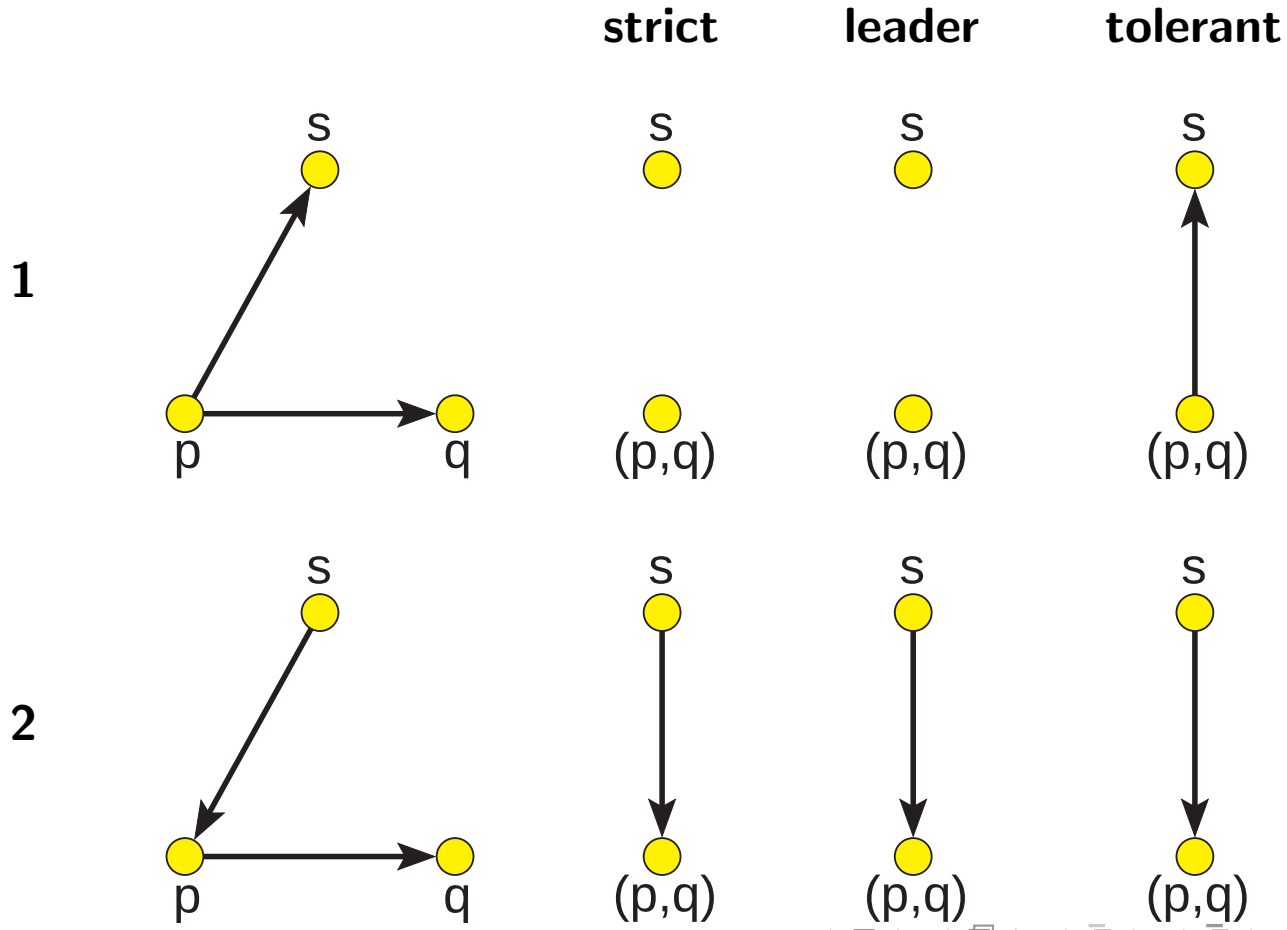
1. $k := n$; $\mathbf{C}(k) := \{\{X\} : X \in \mathbf{U}\}$; $h_D(X) = 0, X \in \mathbf{U}$
2. **while** $\exists C_i, C_j \in \mathbf{C}(k)$: $(i \neq j \wedge \psi(C_i, C_j))$ **repeat**
2.1. $(C_p, C_q) := \text{argmin}\{D(C_i, C_j) : i \neq j \wedge \psi(C_i, C_j)\}$;
2.2. $C := C_p \cup C_q$; $k := k - 1$; $h_D(C) = D(C_p, C_q)$
2.3. $\mathbf{C}(k) := \mathbf{C}(k+1) \setminus \{C_p, C_q\} \cup \{C\}$;
2.4. adjust the relation R as required by the clustering type
2.5. determine $D(C, C_s)$ for all $C_s \in \mathbf{C}(k)$, $\psi(C, C_s)$
3. $m := k$

The *fusibility condition* $\psi(C_i, C_j)$ is equivalent to $C_i R C_j$ for tolerant, leader and strict method; and to $C_i R C_j \wedge C_j R C_i$ for two-way method.

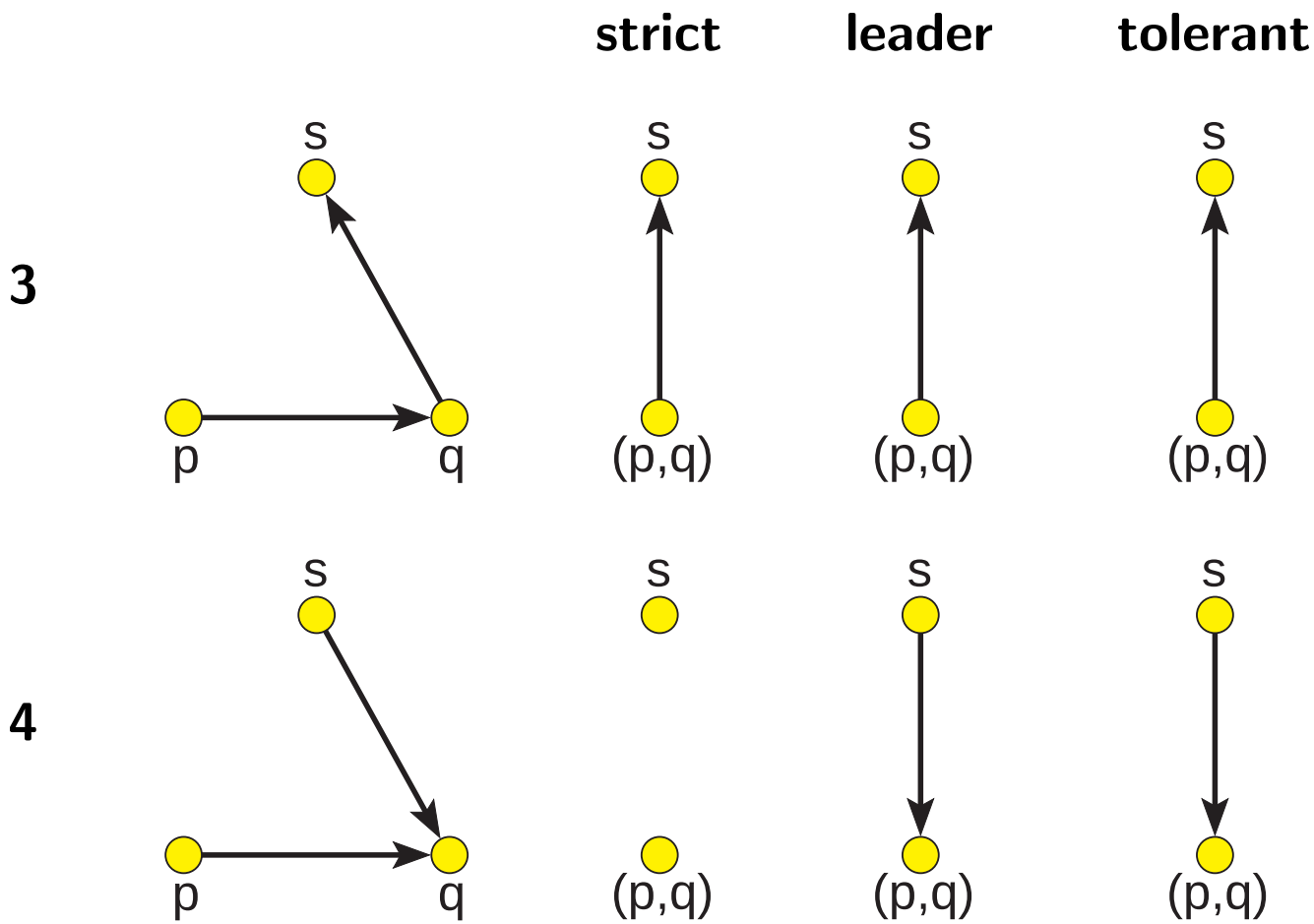
To get clustering procedures, it is necessary to further elaborate the questions how to adjust the relation after joining two clusters and how to update the dissimilarity $D(C, C_s)$.

Types of relational constraints

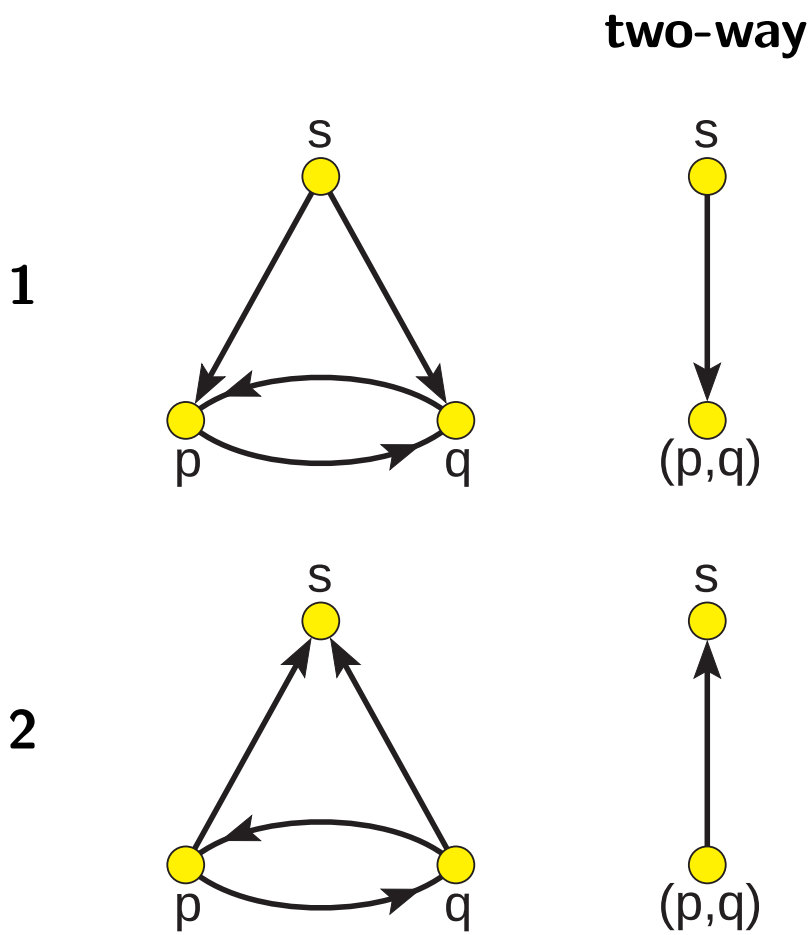
In figures four adjusting *strategies* are presented. They are compatible with the corresponding types of constraints: Φ^1 – tolerant, Φ^2 – leader, Φ^4 – strict, and Φ^5 – two.way.



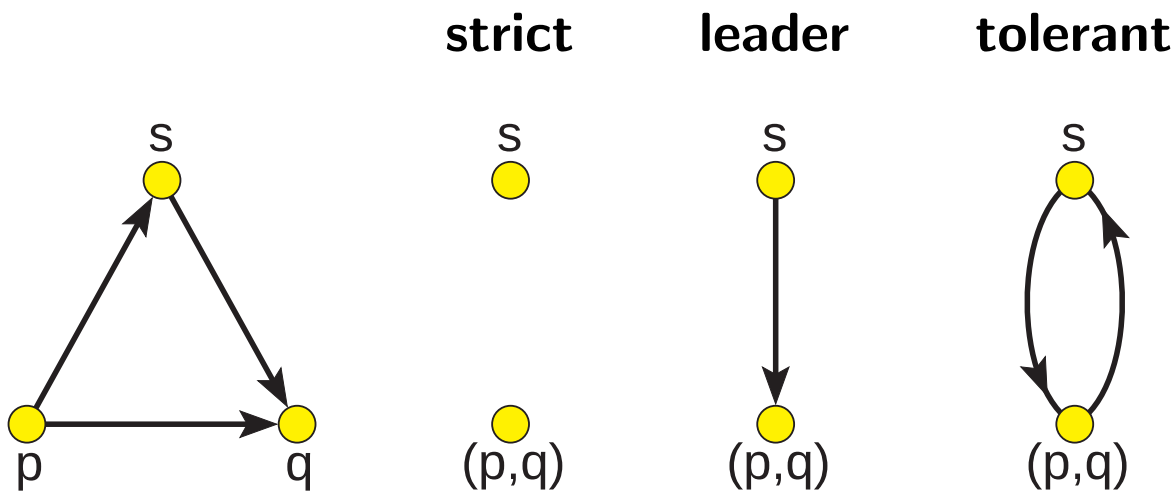
Types of relational constraints



The two-way strategy



An example of application of strategies



Adjusting dissimilarity

In the original approach (Ferligoj 1983), a complete dissimilarity matrix is needed. To obtain fast algorithms that can be applied to large data sets we propose *considering only the dissimilarities between linked units*. For large data sets, we assume that the relation R is *sparse*.

For step 2.4, "determine $D(C, C_s)$ for all $C_s \in \mathbf{C}(k)$" in the agglomerative procedure requires the adjustment of dissimilarities – computing the dissimilarities between a new cluster C and other remaining clusters. In the case of the relational constraints, we can limit the computation only to clusters that are related/linked to C.

This can be done efficiently in the following two ways:

A: we define a dissimilarity $D(S, T)$ between clusters S and T that allows quick updates (as in Lance-Williams formula).

B: to each cluster we assign a representative and can efficiently compute a representative of merged clusters along with a dissimilarity between clusters in terms of their representatives.

The first approach

We will present only he first approach. Let $(\mathbf{U}, R)$, $R \subseteq \mathbf{U} \times \mathbf{U}$ be a graph and $\emptyset \subset S, T \subset \mathbf{U}$ and $S \cap T = \emptyset$. We call a *block* of relation R for S and T its part $R(S, T) = R \cap S \times T$. The *symmetric closure* of relation R we denote with $\hat{R} = R \cup R^{-1}$. It holds: $\hat{R}(S, T) = \hat{R}(T, S)$.
For all dissimilarities between clusters $D(S, T)$ we set:

$$D(\{s\}, \{t\}) = \begin{cases} d(s,t) & s \,\hat{R}\, t \\ \infty & \text{otherwise} \end{cases}$$

where d is a selected dissimilarity between units.
Minimum

$$D_{\min}(S, T) = \min_{(s,t) \in \hat{R}(S,T)} d(s,t)$$

$$D_{\min}(S, T_1 \cup T_2) = \min(D_{\min}(S, T_1), D_{\min}(S, T_2))$$

The first approach

Maximum

$$D_{\max}(S,T) = \max_{(s,t)\in\hat{R}(S,T)} d(s,t)$$

$$D_{\max}(S, T_1 \cup T_2) = \max(D_{\max}(S, T_1), D_{\max}(S, T_2))$$

Average

$w : V \to \mathbb{R}$ – is a weight on units; for example $w(v) = 1$, for all $v \in \mathbf{U}$.

$$D_{\mathrm{a}}(S,T) = \frac{1}{w(\hat{R}(S,T))} \sum_{(s,t)\in\hat{R}(S,T)} d(s,t)$$

$$w(\hat{R}(S, T_1 \cup T_2)) = w(\hat{R}(S, T_1)) + w(\hat{R}(S, T_2))$$

$$D_{\mathrm{a}}(S, T_1 \cup T_2) = \frac{w(\hat{R}(S, T_1))}{w(\hat{R}(S, T_1 \cup T_2))} D_{\mathrm{a}}(S, T_1) + \frac{w(\hat{R}(S, T_2))}{w(\hat{R}(S, T_1 \cup T_2))} D_{\mathrm{a}}(S, T_2)$$

Reducibility

The dissimilarity D has the *reducibility* property (Bruynooghe, 1977) iff for all C_p, C_q and C_s

$$D(C_p, C_q) \leq \min(D(C_p, C_s), D(C_q, C_s)) \Rightarrow$$

$$\min(D(C_p, C_s), d(C_q, C_s)) \leq D(C_p \cup C_q, C_s)$$

Theorem: If a dissimilarity D has the reducibility property then h_D is a level function.

Theorem: Dissimilarities D_{min}, D_{max} and D_a have the reducibility property.

All three dissimilarities have the reducibility property. In this case, also the *nearest neighbor network* for a given network is preserved after joining the nearest clusters. This allows us to develop a very fast agglomerative hierarchical clustering procedure [Murtagh(1985)] and [Batagelj et al.(2014)] (Subsection 9.3.5). It is available in the program `Pajek`. The same approach can be extended also to clustering of links of network [Bodlaj and Batagelj(2015)] by transforming a given network into its line-graph in which the original links become new nodes.

Example: Citations among authors from the network clustering literature

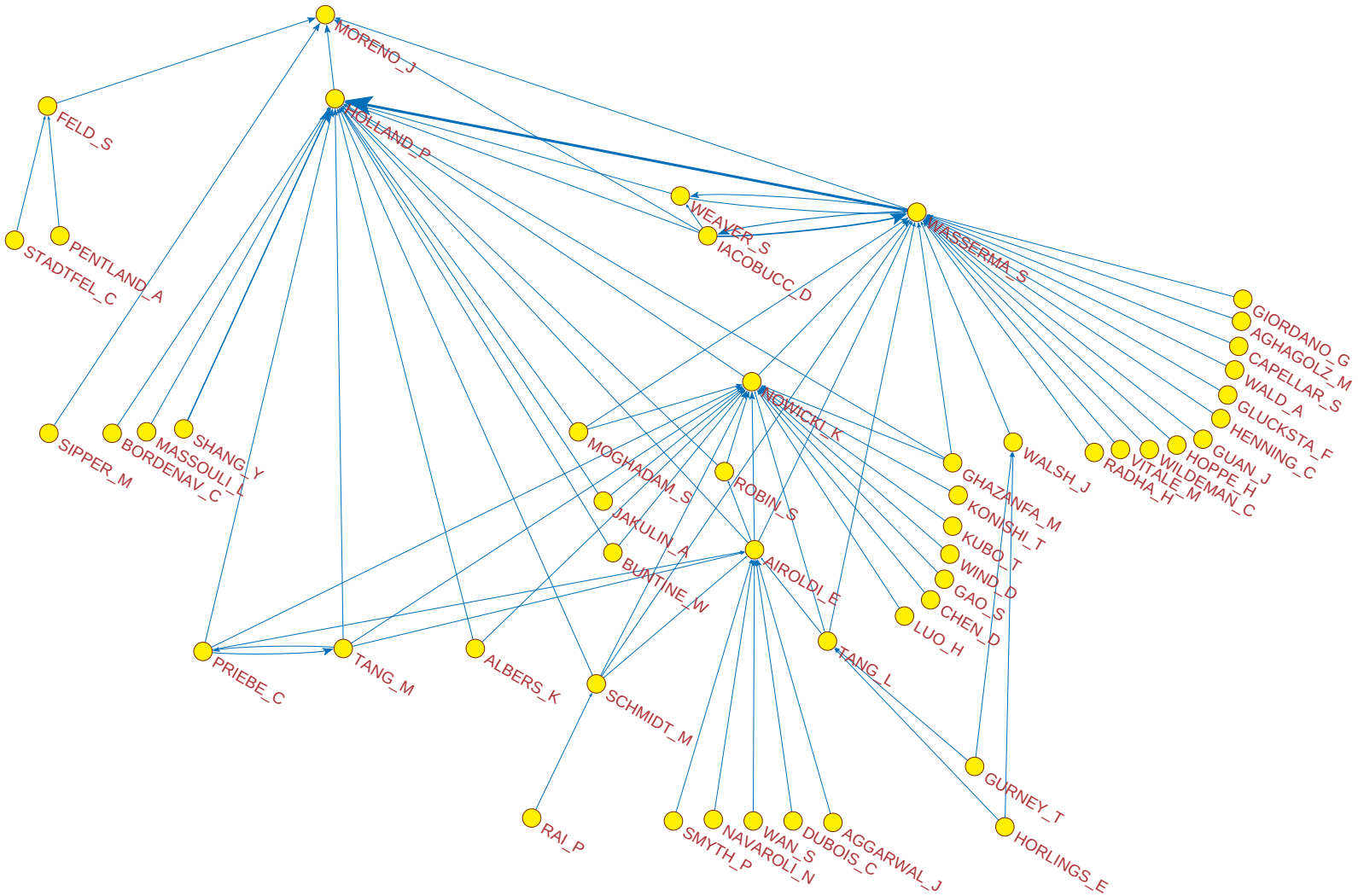
We consider the bibliometric data on the network clustering literature. We analyze the normalized network of citations among authors $\mathbf{nAcite} = n(\mathbf{WAc})^T * n(\mathbf{CiteC}) * n(\mathbf{WAc})$. The weight $\mathbf{nAcite}[u, v]$ of the arc (u, v) is equal to the fractional share of works co-authored by u that are citing a work co-authored by v.
$W = 5695$, $A = 13376$

We identified clusters such that the corresponding induced subnetworks are connected and contain a single center – type Φ^2. The **nAcite** weights are similarities, $s \in [\infty, 0]$. To convert them to dissimilarities d we used the transformation $d = 1 - \frac{s}{s_{max}} \in [0, 1]$, $s_{max} = 2.52$.

On the obtained network, we applied, in Pajek, the hierarchical clustering with relational constraints procedure with the Maximum/Leader strategy and determined the partition of units into clusters of size at most 50. There are 257 such clusters. To reduce their number, we decided to consider only clusters with at least 20 units. There are 57 such clusters. Most of the subnetworks of clusters for the Leader strategy have almost acyclic structure. This has to be considered also in their visualization.

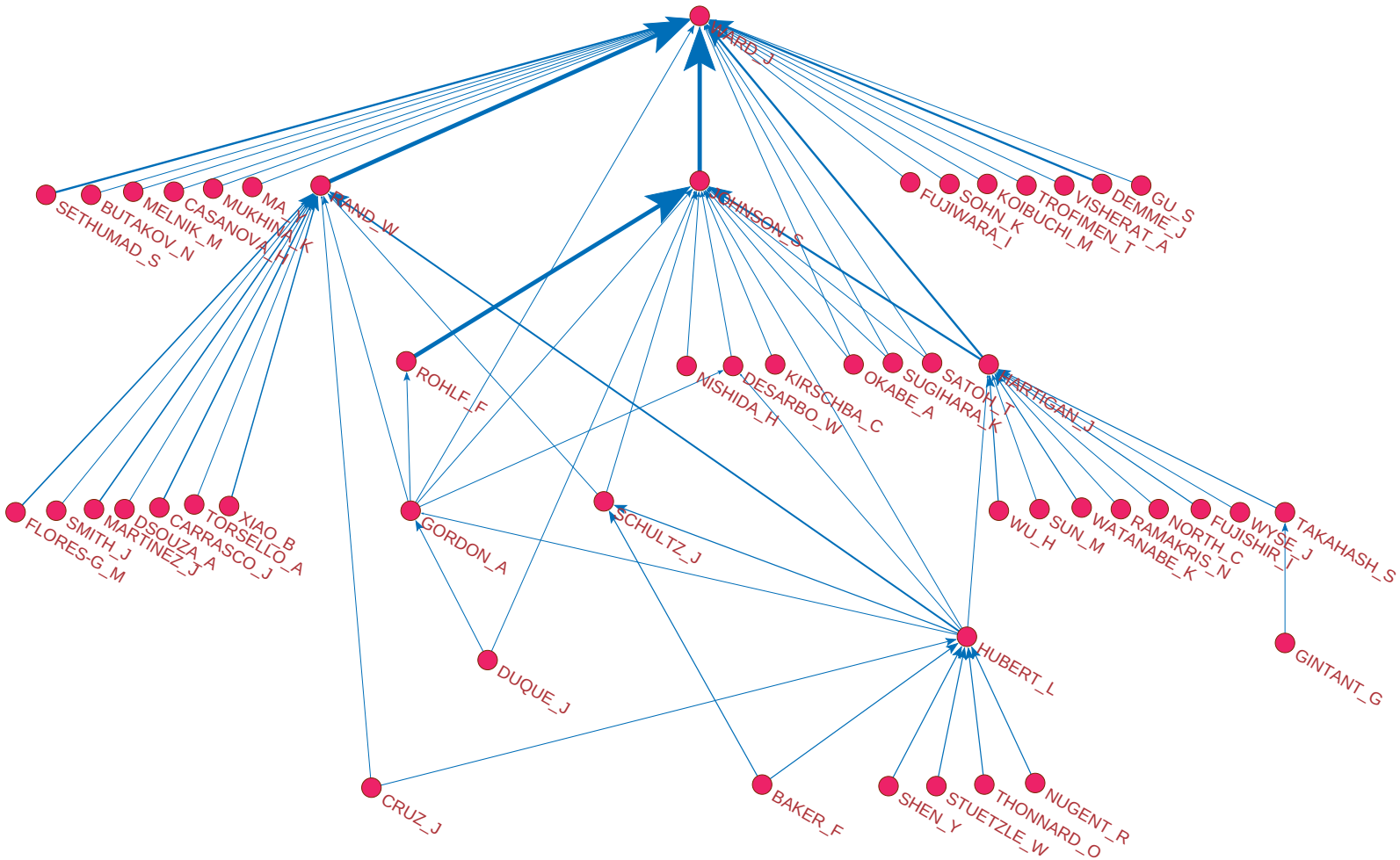
Citations in network clustering literature

Moreno / Wasserman's subnetwork



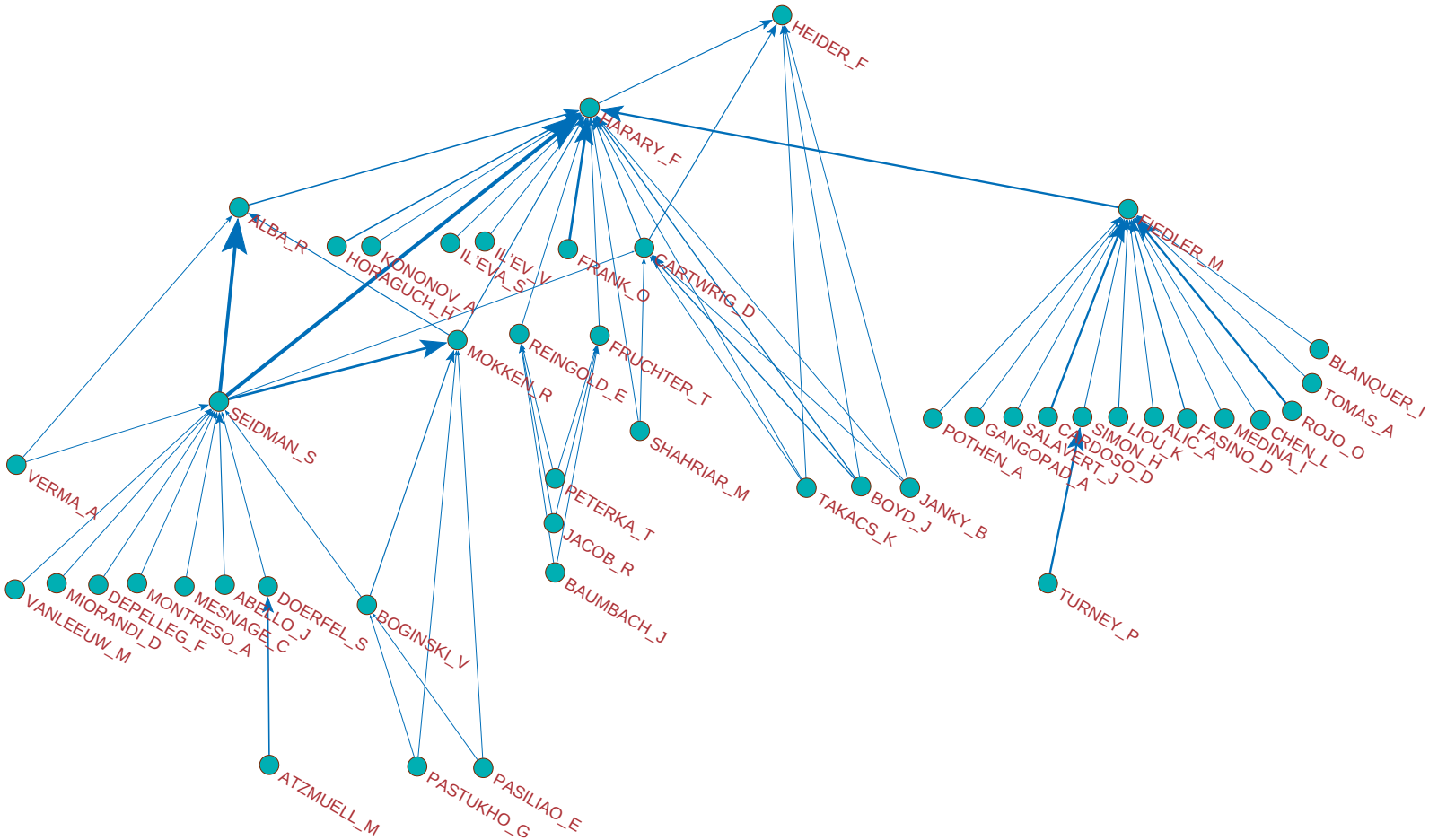
Citations in network clustering literature

Ward's subnetwork



Citations in network clustering literature

Heider / Harary's subnetwork



Citations in network clustering literature

Lefkovitz / Batagelj + Ferligoj's subnetwork

Clustering in bibliometric networks

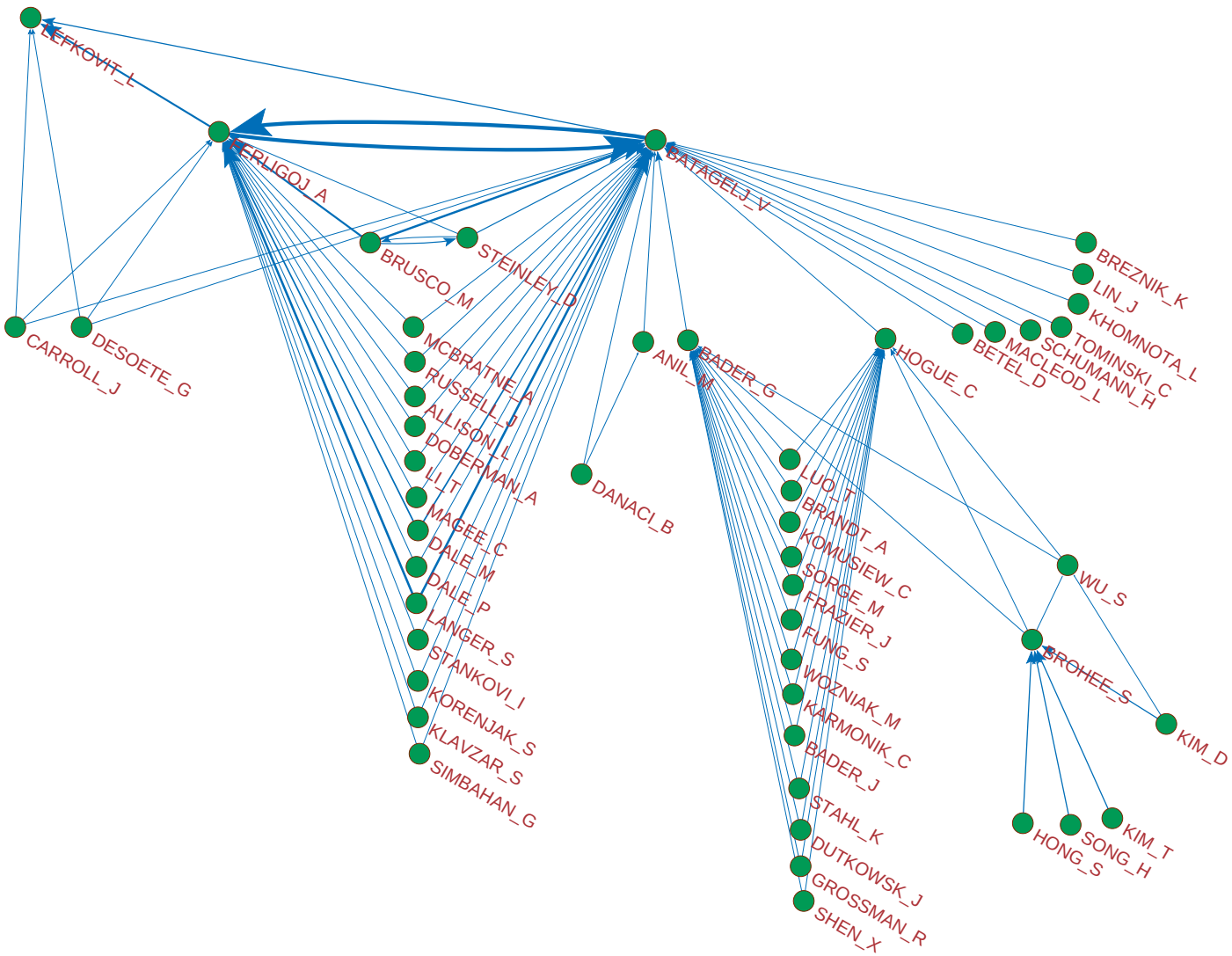
V. Batagelj

Networks

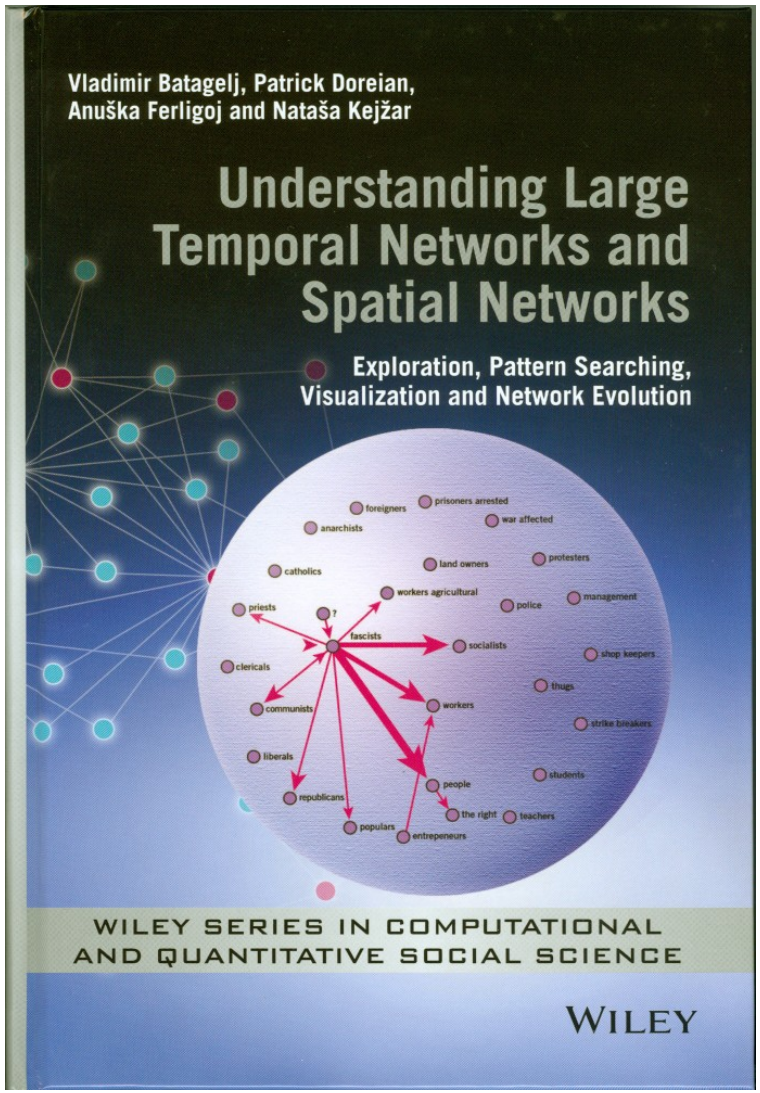
Clustering with relational constraint

Example: Leader strategy

References



Understanding large networks



For details on analysis of large networks see chapters 2 and 3 in the book:

V. Batagelj, P. Doreian, A. Ferligoj and N. Kejžar: Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science. Wiley, October 2014.

References I

M. R. Anderberg. *Cluster Analysis for Application*. Academic Press, New York, 1973.

V. Batagelj. Generalized Ward and related clustering problems. In H. Bock, editor, *Classification and Related Methods of Data Analysis*, pages 67–74. North-Holland, Amsterdam, 1988.

V. Batagelj. Similarity measures between structured objects. In A. Graovac, editor, *MATH/CHEM/COMP 1988: proceedings of an International Course and Conference on the Interfaces between Mathematics, Chemistry, and Computer Science, Dubrovnik, Yugoslavia, 20-25 June 1988*, Studies in physical and theoretical chemistry, pages 25–40. Elsevier, 1989.

Batagelj, V.: Wos2pajek – networks from web of science (2007). http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek

V. Batagelj. **clurc** – R package for clustering with relational constraint. 2017. URL `https://github.com/bavla/cluRC`.

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845-864.

References II


V. Batagelj and A. Ferligoj. Clustering relational data. In W. Gaul, O. Opitz, and M. Schader, editors, *Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 3–15. Springer, Berlin, Heidelberg, 2000.


V. Batagelj, P. Doreian, A. Ferligoj, and N. Kejžar. *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution*. Wiley Series in Computational and Quantitative Social Science Series. Wiley, 2014.


J. Bodlaj and V. Batagelj. Hierarchical link clustering algorithm in networks. *Physical Review E*, 91(6):062814, 2015.


M. Bruynooghe. Méthodes nouvelles en classification automatique des données taxinomiques nombreuses. *Statistique et Analyse des Données*, (3): 24–42, 1977.


A. Ferligoj and V. Batagelj. Some types of clustering with relational constraints. *Psychometrika*, 48(4):541–552, 1983.


A. Ferligoj and V. Batagelj. Clustering with relational constraint. *Psychometrika*, 47(4):413–426, 1982.

References III


J. A. Hartigan. *Clustering algorithms*. Wiley-Interscience, New York, 1975.


L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. A Wiley-Interscience publication. Wiley, 1990.


F. Murtagh. *Multidimensional clustering algorithms*, volume 4. Physika Verlag, Vienna, 1985.


W. D. Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek, 3rd edition*. Cambridge University Press, New York, NY, USA, 2018.


Zaveršnik, M., Batagelj, V.: Islands. In: XXIV International Sunbelt Social Network Conference, Portorož, Slovenia (2004)


Pajek's wiki. http://pajek.imfm.si


Vladimir Batagelj, Andrej Mrvar: Pajek manual.

Acknowledgments

This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J1-9187 and J7-8279) and by Russian Academic Excellence Project '5-100'.